AFOSR-TR- 85-0015

INFORMATIVE GEOMETRY OF PROBABILITY SPACES

Jacob Burbea*

Center for Multivariate Analysis

University of Pittsburgh

INFORMATIVE GEOMETRY OF PROBABILITY SPACES

Jacob Burbea*

December 1984

Technical Report No. 84-52

Center for Multivariate Analysis
515 Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

*Part of the work of this author is sponsored by the Air Force Office of Scientific Research (AFSC), under Contract F49620-85-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation heron.

INFORMATIVE GEOMETRY OF PROBABILITY SPACES

Jacob Burbea

ABSTRACT

The paper is concerned with the geometrical properties that are induced by the local information contents and structures of the parameter space of probability distributions. Of particular interest in this investigation is the Rao distance which is the geodesic distance induced by the differential metric associated with the Fisher information matrix of the parameter space. Moreover, following Efron, Dawid and Amari, some affine connections are introduced into the informative geometry of parameter space and thereby elucidating the role of the curvature in statistical studies. In addition, closed form expressions of the Rao distances for certain families of probability distributions are given and discussed.

Informative Geometry of Probability Spaces

by Jacob Burbea

Metrics and distances (or semi-distances) between probability distributions play an important role in problems of statistical inference and in pratical applications to study affinities among a given set of populations. A statistical model is specified by a family of probability distributions, usually described by a set of continuous parameters known as parameter space. The latter possesses some geometrical properties which are induced by the local information contents and structures of the distributions. Starting from Fisher's pioneering work [17] in 1925, the study of these geometrical properties has received much attention in the statistical literature. In 1945, Rao [24] introduced a Riemannian metric in terms of the Fisher information matrix over the parameter space of a paremetric family of probability distributions, and proposed the geodesic distance induced by the metric as a measure of dissimilarity between two probability distributions. Since then, many statisticians have attempted to construct a geometrical theory in probability spaces and it was only after thirty years later that Efron [13] was able to introduce a new affine connection into the geometry of parameter spaces and thereby elucidating the important role of the curvature

in statistical studies. Significant contributions to Efron's work were made by Reeds [28] and Dawid [11]. The latter has even suggested a geometrical foundation for Efron's approach as well as pointing out the possibility of introducing other affine connections into the geometry of parameter spaces (see also Amari [1,2]). This recent study has also revived the interest in dissimilarity measures like the Rao distance [25], especially in the closed form expressions of these distances for certain families of probability distributions. Some work in these directions was done earlier by Čenčov [9, 10]. Recently, Atkinson and Mitchell [3], independently of Čenčov [9,10], computed the Rao distances for a number of parametric families of probability distributions. A unified approach to the construction of distance and dissimilarity measures in probability spaces is given in recent papers by Burbea and Rao [7,8], and Oller and Cuadras [22] (see also [6]).

1. Generalities.

We first introduce some notation. Let μ be a σ-finite additive measure, defined on a σ-algebra of the subsets of a measurable space χ. Then, $M \equiv M(\chi:\mu)$ stands for the space of all μ-measurable functions on χ, $L \equiv L(\chi:\mu)$ designates the space of all $p \in M$ so that

$$||p||_\mu \equiv \int_\chi |p(x)|\,d\mu(x) = \int_\chi |p|\,d\mu < \infty.$$

By $M_+ \equiv M_+(\chi:\mu)$ we denote the set of all $p \in M$ such that $p(x) \in R_+ \equiv (0,\infty)$ for μ-almost all $x \in \chi$, and we define $L_+ \equiv L_+(\chi:\mu)$ as $L_+ = M_+ \cap L$. We let $P \equiv P(\chi:\mu)$ stand for the set of all $p \in L_+$ with $||p||_\mu = 1$. Evidently, P is a convex subset of L_+, and $p \in L_+$ if and only if $p/||p||_\mu \in P$.

In the probability context, a random variable X takes values in the sample space χ according to a probability distribution p assumed to belong to P. If X is a continuous random variable, μ will be the Lebesgue measure on the Borel sets of a euclidean sample space χ and, if X is discrete, μ is taken as a counting measure on the sets of a countable sample space χ.

Let $\theta = (\theta_1,\ldots,\theta_n)$ be a set of real continuous parameters belonging to a parameter space Θ, a manifold embedded in R^n and let $F_\Theta = \{p(\cdot|\theta) \in L_+ : \theta \in \Theta\}$ be a parametric family of positive distributions $p = p(\cdot|\theta)$, $\theta \in \Theta$, with some regularity properties not mentioned explicitly to avoid lengthy discussion (see, however [3,7,12]). For example, it is implicitly assumed that

$$\partial_i p \equiv \partial_i p(\cdot|\theta) \equiv \partial p(\cdot|\theta)/\partial\theta_i \qquad (p = p(\cdot|\theta) \in F_\Theta\ ,\ \ i = 1,\ldots,n)$$

is in M for every $\theta \in \Theta$. It is also assumed that for a fixed $0 \in \Theta$, the n functions $\{\partial_i p\}_{i=1}^n$ are linearly independent over

χ. We also consider a parametric family of probability distributions $P_\Theta=\{p(\cdot|\theta)\varepsilon P:\theta\varepsilon\Theta\}$ which may be viewed as a convex subfamily of F_Θ.

Let f be a continuous and positive function on $\mathbb{R}_+$ and define

$$ds_f^2(\theta)\equiv\int_\chi \frac{f(p)}{p}[dp]^2 d\mu \qquad (\theta\varepsilon\Theta\ ,\ p=p(\cdot|\theta)\varepsilon F_\Theta),$$

where in the integrand, the dependence on $x\varepsilon\chi$ and $\theta\varepsilon\Theta$ is supressed and where

$$dp=dp(\cdot|\theta)=\sum_{i=1}^{n}(\partial_i p)d\theta_i.$$

Here and throughout the remaining parts of this entry we shall use freely the convention of supressing the dependence on $x\varepsilon\chi$ and $\theta\varepsilon\Theta$. Thus, with this convention,

$$ds_f^2(\theta)=\sum_{i,j=1}^{n} g_{ij}^{(f)}d\theta_i d\theta_j \tag{1.1}$$

with

$$g_{ij}^{(f)}=g_{ij}^{(f)}(\theta)=\int_\chi \frac{f(p)}{p}(\partial_i p)(\partial_j p)d\mu.$$

It follows that the $n\times n$ matrix $[g_{ij}^{(f)}(\theta)]$ is positive-definite for every $\theta\varepsilon\Theta$ and hence ds_f^2 gives a Riemannian metric on Θ. Alternative expressions for these quantities are available through the language of expectations. Thus, for $p=p(.|\theta)\ \varepsilon\ F_\Theta$,

$$ds_f^2(\theta)=E_\theta[(fop)(dlogp)^2]$$

and

$$g_{ij}^{(f)}(\theta)=E_\theta[(f\circ p)(\partial_i \log p)(\partial_j \log p)].$$

In the theory of information (see [6]) the quantity $-\log p(\cdot|\theta)$, for $p(\cdot|\theta)\in P_\Theta$, is known there as the amount of "self-information" associated with the state $\theta\in\Theta$. The self-information for the nearby state $\theta+\delta\theta\in\Theta$ is then $-\log p(\cdot|\theta+\delta\theta)$. To the first order, the difference between the self-informations associated with these states is given by

$$d\log p=\sum_{i=1}^{n}(\partial_i \log p)d\theta_i$$

and hence $ds_f^2(\theta)$ measures the weighted average of the square of this first order difference with the weight $f[p(\cdot|\theta)]$. For this reason, the metric ds_f^2 and the matrix $[g_{ij}^{(f)}]$ are called the "f-information metric" and the "f-information matrix", respectively.

As is well-known from differential geometry, $g_{ij}^{(f)}$ $(i,j=1,\ldots,n)$ is a convariant symmetric tensor of the second order for all $\theta\in\Theta$, and hence ds_f^2 is invariant under the admissible transformations of the paremeters. Let $\theta=\theta(t)$, $t_1\leq t\leq t_2$, be a curve in Θ joining the points $\theta^{(1)}$, $\theta^{(2)}\in\Theta$ with $\theta^{(j)}=\theta(t_j)(j=1,2)$. Since $ds_f\equiv(ds_f^2)^{1/2}$ is the line element of the metric ds_f^2, the distance between these points

along this curve is

$$\left|\int_{t_1}^{t_2} \frac{ds_f}{dt}\, dt\right| = \left|\int_{t_1}^{t_2} \{\sum_{i,j=1}^{n} g_{ij}^{(f)}(\theta)\dot{\theta}_i\dot{\theta}_j\}^{1/2} dt\right|$$

where a dot denotes differentiation with respect to the curve-parameter t. The geodesic curve, namely the curve joining $\theta^{(1)}$ and $\theta^{(2)}$ such that the above distance is the shortest is called the "f-information geodesic curve" along $\theta^{(1)}$ and $\theta^{(2)}$ while the resulting distance $S_f(\theta^{(1)},\theta^{(2)})$ is called the "f-information geodesic distance" between $\theta^{(1)}$ and $\theta^{(2)}$. The f-information geodesic curve $\theta=\theta(t)$ may be determined from the Euler-Lagrange equations

$$(1.2) \qquad \sum_{i=1}^{n} g_{ik}^{(f)}\ddot{\theta}_i + \sum_{i,j=1}^{n} \Gamma_{ijk}^{(f)}\dot{\theta}_i\dot{\theta}_j = 0 \qquad (k=k,\ldots,n)$$

and from the boundary conditions

$$\theta_i(t_j)=\theta_i^{(j)} \qquad (i=1,\ldots,n\ ;\ j=1,2).$$

Here, the quantity $\Gamma_{ijk}^{(f)}$ is given by

$$(1.3) \qquad \Gamma_{ijk}^{(f)} = \frac{1}{2}[\partial_i g_{jk}^{(f)} + \partial_j g_{ki}^{(f)} - \partial_k g_{ij}^{(f)}]$$

and is known as the "Christoffel symbol of the first kind" for the metric ds_f^2.

By the very definition of the f-information geodesic curve $\theta=\theta(t)$, its tangent vector $\dot{\theta}=\dot{\theta}(t)$ is of constant length with respect to the metric ds_f^2. Thus,

$$(1.4)\qquad \{\dot{s}_f(\theta(t))\}^2=\sum_{i,j=1}^{n} g_{ij}^{(f)}\dot{\theta}_i\dot{\theta}_j\equiv\text{const.}$$

The constant may be chosen to be of value 1 when the curve-parameter t is chosen to be the arc-length parameter s, $0\leq s\leq s_o$ with $s_o\equiv S_f(\theta^{(1)},\theta^{(2)})$, $\theta(0)=\theta^{(1)}$ and $\theta(s_o)=\theta^{(2)}$. It is also clear that the f-information geodesic distance S_f on the parameter space Θ is invariant under the admissible transformations of the parameters as well as of the random variables.

The metric $ds_f^2(\theta)$ may also be regarded as a functional of $p(\cdot|\theta)\in F_\Theta$. This functional is convex in $p(\cdot|\theta)\in F_\Theta$ if and only if the function $F(x)\equiv x/f(x)$ is concave on $\mathbb{R}_+$. In particular, if f is also a C^2-function on $\mathbb{R}_+$ then this holds if and only if $FF''\geq 2(F')^2$ on $\mathbb{R}_+$. The choice of $f(x)=x^{\alpha-1}$ gives the "<u>α-order information metric</u>"

$$(1.5)\qquad ds_\alpha^2(\theta)=E_\theta[p^{\alpha-1}(d\log p)^2]$$

with the corresponding "<u>α-order information matrix</u>"

$$(1.6)\qquad g_{ij}^{(\alpha)}(\theta)=E_\theta[p^{\alpha-1}(\partial_i\log p)(\partial_j\log p)]$$

and the "<u>α-order information geodesic distance</u>" S_α on Θ. It follows that $ds_\alpha^2(\theta)$ is convex in $p(\cdot|\theta)\in F_\Theta$ if and only if $1\leq\alpha\leq 2$. We drop the suffix α when $\alpha=1$. Then, ds^2 is known as the "<u>information metric</u>" or the "<u>Fisher amount of information</u>" while $[g_{ij}]$ is the well-known "<u>information matrix</u>"

or the "Fisher information matrix". The distance S on Θ is called the "information geodesic distance" or the "Rao distance" (see [3,7,26]). We also note that

$$g_{ij}^{(\alpha)}(\theta)=\int_\chi p^{\alpha-1}\partial_i\partial_j p d\mu-\int_\chi p^\alpha \partial_i\partial_j \log p d\mu.$$

Moreover, for $\alpha\neq 0$,

$$g_{ij}^{(\alpha)}(\theta)=\alpha^{-2}\partial_i\partial_j\int_\chi p^\alpha d\mu-\alpha^{-1}\int_\chi p^\alpha \partial_i\partial_j \log p d\mu.$$

In particular,

$$g_{ij}(\theta)=\partial_i\partial_j\int_\chi p d\mu-\int_\chi p\partial_i\partial_j \log p d\mu$$

and thus

$$(1.7)\quad g_{ij}(\theta)=-\int_\chi p\partial_i\partial_j \log p d\mu=-E_\theta(\partial_i\partial_j \log p) \qquad (p(\cdot|\theta)\in P_\Theta).$$

The metric $ds_f^2(\theta)$ arises as the second order differential of certain entropy or divergence functionals along the direction of the tangent space of Θ at $\theta\in\Theta$. See [7,12] for more details (see also [6]).

For example, let $F(\cdot,\cdot)$ be a C^2-function on $\mathbb{R}_+\times\mathbb{R}_+$ and consider the "F-divergence"

$$D_F(p,q)\equiv\int_\chi F[p(x),q(x)]d\mu(x) \qquad (p,q\in M_+).$$

We shall also assume that F satisfies the following additional properties: (i) $F(x,\cdot)$ is strictly convex on $\mathbb{R}_+$ for every $x\in\mathbb{R}_+$; (ii) $F(x,x)=0$ for every $x\in\mathbb{R}_+$;

(iii) $\partial_y F(x,y)\big|_{y=x}$=const. for every $x\varepsilon R_+$. For $p(\cdot|\theta^{(1)})$ and $p(\cdot|\theta^{(2)})$ in P_Θ we write

$$\mathcal{D}_F(\theta^{(1)},\theta^{(2)})\equiv D_F[p(\cdot|\theta^{(1)}),p(\cdot|\theta^{(2)})] \qquad (\theta^{(1)},\theta^{(2)}\varepsilon\Theta).$$

Then, for $p(\cdot|\theta)\varepsilon F_\Theta$ and $\theta\varepsilon\Theta$,

$$\mathcal{D}_F(\theta,\theta)=0 \ , \ d\mathcal{D}_F(\theta,\theta)=\int_\chi \partial_y F(p,y)\big|_{y=p}(dp)d\mu=0$$

and

$$d^2\mathcal{D}_F(\theta,\theta)=ds_f^2(\theta)$$

where

$$f(x)=x\partial_y^2 F(x,y)\big|_{y=x} \qquad (x\varepsilon R_+).$$

It follows that to the second order infinitesimal displacements

$$\mathcal{D}_F(\theta,\theta+\delta\theta)=\frac{1}{2}ds_f^2(\theta).$$

2. Properties of the Information Metric.

We shall describe some further properties of the f-information metric ds_f^2. However, for reasons of clarity and economy we shall restrict ourselves here to the case of the ordinary information metric ds^2 (i.e. when $f(x)\equiv 1$ or when $\alpha=1$ in (1.5)) on the parametric space Θ of probability distributions $p(\cdot|\theta)$ in P_Θ. A more general discussion may be found in [7,8]. We shall hereafter also assume that the sum-

mation is taken without the symbol Σ when the indices are repeated twice and that the extent of the summation is understood as running from 1 to n. Thus, with this convention, we have, by virtue of (1.1), (1.3), (1.6) and (1.7),

$$ds^2=g_{ij}d\theta_i d\theta_j$$

$$g_{ij}=E_\theta[(\partial_i \log p)(\partial_j \log p)]=-E_\theta[\partial_i\partial_j \log p]$$

and

$$\Gamma_{ijk}=\frac{1}{2}[\partial_i g_{jk}+\partial_j g_{ki}-\partial_k g_{ij}]. \tag{2.1}$$

The information geodesic curves $\theta=\theta(s)$, where s is the arc-length parameter, are determined, in view of (1.2), by

$$g_{ij}\ddot{\theta}_i+\Gamma_{ijk}\dot{\theta}_i\dot{\theta}_j=0 \qquad (k=1,\ldots,n). \tag{2.2}$$

Moreover, from (1.4) we also have

$$g_{ij}\dot{\theta}_i\dot{\theta}_k=1. \tag{2.3}$$

Thus, for two points $a,b\in\Theta$, or for $p(\cdot|a)$, $p(\cdot|b)\in P_\Theta$, the Rao distance $S(a,b)$ is completely determined by (2.2), a system of n second order (non-linear) differenital equations, and by the 2n boundary conditions $\theta(0)=a$ and $\theta(s_o)=b$ with $s_o=S(a,b)$. This computation may be facilitated with the aid of the normalization (2.3).

We denote by I the Fisher information matrix $[g_{ij}]$, by

g^{ij} the elements of its inverse I^{-1}, and, as usual, the elements of the unit matrix I are denoted by the delta of Kronecker δ_{ij}. Note that $I^{-1}=[g^{ij}]$ is also positive-definite and that I is associated with a distribution $p(\cdot|\theta)\varepsilon P_\theta$ of a random variable X. We list the following properties (see Rao [27, p. 323-332] for more details):

1° Let I_1 and I_2 be the information matrices due to two independent random variables X_1 and X_2. Then $I=I_1+I_2$ is the information matrix due to $X=(X_1, X_2)$ jointly.

2° Let I_T be the information matrix due to a function T of X. Then $I-I_T$ is semi positive-definite.

3° Let $p(\cdot|\theta)\varepsilon P_\theta$ with the corresponding information matrix I. Assume that $\underline{f}=(f_1,\ldots,f_m)$ is a vector of m statistics (random variables) and define $\underline{g}(\theta)=(g_1(\theta),\ldots,g_m(\theta))$ by $g_i(\theta)=E_\theta(f_i)$ $(i=1,\ldots,m)$, i.e. $\underline{f}$ is an unbiased estimator of $\underline{g}(\theta)$. Consider the m×m and m×n matrices $V=[V_{ij}]$ and $U=[V_{ij}]$ given by $V_{ij}=E_\theta[(f_i-g_i)(f_j-g_j)]$ $(i,j=1,\ldots,m)$ and $U_{ij}=E_\theta[f_i\partial_j\log p]$ $(i=1,\ldots,m;\ j=1,\ldots,n)$. Then:

(i) The m×m matrix $V-UI^{-1}U'$ is semi positive-definite for every $\theta\varepsilon\Theta$. The matrix is zero at some $\theta\varepsilon\Theta$ if and only if $\underline{f}=(f_1,\ldots,f_m)$ is of the form $f_i=\lambda_{ik}\partial_k\log p+E_\theta(f_i)$ $(i=1,\ldots,m)$;

(ii) Suppose, in addition, that $\partial_j\int_X f_i(x)p(x|\theta)d\mu(x)=\int_X f_i(x)\partial_j p(x|\theta)d\mu(x)$ $(i=1,\ldots,m;\ j=1,\ldots,n)$. Then U

is the Jacobian-matrix $[\partial_j g_i]$ of $\underline{g}=(g_1,\ldots,g_m)$ with respect to $\theta=(\theta_1,\ldots,\theta_n)$. In particular, when $m=n$ and $\underline{g}(\theta)=\theta$, i.e. $\underline{f}$ is an unbiased estimator of θ, then $V-I^{-1}$ is semi positive-definite.

The last property constitutes the celebrated "Cramer-Rao lower bound theorem", namely that for any unbiased estimator of θ, its covariance matrix dominates the inverse of the Fisher information matrix.

3. Information Connections and Curvatures.

The information metric renders the parameter space Θ as a Riemannian manifold with the metric tensor g_{ij} associated with the distribution $p(\cdot|\theta)\varepsilon P_\Theta$. In this context, the Christoffel symbol of the first kind Γ_{ijk} in (2.1) is called the "first information connection". As is well known from differential geometry, this natural affine connection induces a parallelism on Θ, known as the "Levi-Civita parallelism", which is compatible with the metric tensor g_{ij}, in the sense that the covariant differentiation of the latter vanishes for this connection. Using the summation convention, one introduces the "Christoffel symbol of the second kind" Γ^k_{ij} by

$$\Gamma^k_{ij}=\Gamma_{ijm}g^{mk}. \tag{3.1}$$

This is also called the "second information connection".

With the aid of this connection, the equation for the information geodesic curves (2.2) assumes the alternative form

$$\ddot{\theta}_k+\Gamma^k_{ij}\dot{\theta}_i\dot{\theta}_j=0 \qquad (k=1,\ldots,n). \tag{3.2}$$

In differential geometry one also considers the "Riemann-Christoffel tensor of the second kind"

$$R^{\ell}_{ijk}=\partial_j\Gamma^{\ell}_{ik}-\partial_k\Gamma^{\ell}_{ij}+\Gamma^m_{ik}\Gamma^{\ell}_{mj}-\Gamma^m_{ij}\Gamma^{\ell}_{mk} \tag{3.3}$$

and the "Riemann-Christoffel tensor of the first kind"

$$R_{ijk\ell}=R^m_{jk\ell}g_{mi}. \tag{3.4}$$

These quantities are also known as the "second information curvature tensor" and the "first information curvature tensor", respectively. In this respect, it is worthwhile noticing that

$$R_{ijk\ell}=-R_{jik\ell}=-R_{ij\ell k}=R_{k\ell ij},$$

$$R_{ijk\ell}+R_{ik\ell j}+R_{i\ell jk}=0$$

and that the number of distinct nonvanishing components of the tensor $R_{ijk\ell}$ is $n^2(n^2-1)/12$. The latter reduces to 0 when n=1 and to 1 when n=2.

The "mean Gaussian curvature", in the directions of $x=(x_1,\ldots,x_n)$ and $y=(y_1,\ldots,y_n)$ of $\mathbb{R}^n$ is given by

$$(3.5)\quad \kappa \equiv \kappa(\theta:x,y) = \frac{R_{ijk\ell}x_i y_j x_k y_\ell}{(g_{ik}g_{j\ell}-g_{i\ell}g_{jk})x_i y_j x_j y_\ell} \qquad (\theta\varepsilon\Theta),$$

and is also called the "information-curvature" in the directions of x and y. This curvature is identically zero if Θ is euclidean and is constant if the space Θ is isotropic (i.e. when κ is independent of the directions x and y), provided $n>2$.

Besides the first information connection Γ_{ijk} there are, of course, other connections leading to parallelisms which differ from the Levi-Civita parallelism. However, in the context of statistical inference, the choice of such connections should reflect the structure of the distributions in some meaningful manner. Following an idea of Dawid [11], Amari [1] considers the one parameter family of affine connections $\overset{\alpha}{\Gamma}_{ijk}$ given by

$$\overset{\alpha}{\Gamma}_{ijk} \equiv \Gamma_{ijk} - \frac{\alpha}{2}T_{ijk} \qquad (\alpha\varepsilon R)$$

where T_{ijk} is the symmetric tensor

$$T_{ijk} \equiv E_\theta[(\partial_i \log p)(\partial_j \log p)(\partial_k \log p)].$$

The connection $\overset{\alpha}{\Gamma}_{ijk}$ is called the "α-connection". Thus, in this context, the first information connection is the 0-connection. An alternative expression for the α-connection is

$$\overset{\alpha}{\Gamma}_{ijk} = E_\theta[(\partial_i \partial_j \log p)(\partial_k \log p)] + \frac{1-\alpha}{2}T_{ijk}.$$

The 1-connection was introduced first by Efron [13] and hence is also called the "Efron-connection". The -1-connection, on the other hand, is called the "Dawid-connection", after Dawid [11] who was first to suggest its introduction.

In order to elucidate the relevance and the meaningfulness of the α-connections in statistical problems, we consider two examples suggested by Dawid [11] and described in Amari [1].

Example 1. We consider an exponential family P_Θ of distributions $p(\cdot|\theta)$ given (using the summation convention) by

$$p(x|\theta)=\exp\{T(x)+T_i(x)\theta_i-\psi(\theta)\} \qquad (x\in\chi) \tag{3.6}$$

with

$$e^{\psi(\theta)}=\int_\chi e^{T_i(x)\theta_i}e^{T(x)}d\mu(x) \qquad (\theta\in\Theta), \tag{3.7}$$

and specified by the natural free parameters $\theta=(\theta_1,\ldots,\theta_n)\in\Theta$. Here ψ is a C^2-funtion on Θ, and T and $T_1,\ldots,T_n$ are measurable functions on χ. Under these circumstances, we have

$$\partial_i\log p=T_i(x)-\partial_i\psi(\theta)\ ,\ \partial_i\partial_j\log p=-\partial_i\partial_j\psi.$$

Therefore

$$g_{ij}=\partial_i\partial_j\psi, \tag{3.8}$$

$$E_\theta[(\partial_i\partial_j\log p)(\partial_k\log p)]=0 \tag{3.9}$$

and

$$\overset{\alpha}{\Gamma}_{ijk}=\frac{1-\alpha}{2}T_{ijk}. \tag{3.10}$$

Since $\overset{\alpha}{\Gamma}_{ijk}(\theta)$ is identically zero for $\alpha=1$, we find that the exponential family constitutes an uncurved space with respect to the Efron-connection. For this reason, the Efron-connection may also be called the "exponential-connection".

Example 2. We consdier a family $P_\Theta \equiv P_\Theta(q_1,\ldots,q_{n+1})$ of distributions $p(\cdot|\theta)$ given by a mixture of n+1 prescribed linearly independent probability distributions on χ,

$$p(x|\theta)=q_i(x)\theta_i+q_{n+1}(x)\theta_{n+1} \qquad (x\in\chi)$$

where

$$\theta_{n+1}\equiv 1-(\theta_1+\cdots+\theta_n)$$

and $\theta\in\Theta$ with

$$\Theta=\{\theta=(\theta_1,\ldots,\theta_n)\in R_+^n : \theta_{n+1}>0\}.$$

In this case, we have

$$\partial_i \log p=p^{-1}(q_i-q_{n+1}) \ , \ \partial_i\partial_j \log p=-(\partial_i \log p)(\partial_j \log p).$$

Therefore

$$E_\theta[(\partial_i\partial_j \log p)(\partial_k \log p)]=-T_{ijk}$$

and

$$\overset{\alpha}{\Gamma}_{ijk}=-\frac{1+\alpha}{2}\Gamma_{ijk}.$$

It follows, since $\overset{\alpha}{\Gamma}_{ijk}(\theta)$ is identically zero for $\alpha=-1$, that this family of mixture distributions constitutes an uncurved space with respect to the Dawid-connection. For this reason, the Dawid-connection is also called the "mixture-connection".

Once the α-connection $\overset{\alpha}{\Gamma}_{ijk}$ is adopted, the other related quantities $\overset{\alpha k}{\Gamma}_{ij}$, $\overset{\alpha}{R}{}^{\ell}_{ijk}$, $\overset{\alpha}{R}_{ijk\ell}$ and $\overset{\alpha}{\kappa}$ are determined by the same rules, (3.1) and (3.3)-(3.5), for determining the corresponding quantities when $\alpha=0$. For example,

$$\overset{\alpha k}{\Gamma}_{ij}=\overset{\alpha}{\Gamma}_{ijm}g^{mk}$$

and, corresponding to (3.2), the equation

$$\ddot{\theta}_k+\overset{\alpha k}{\Gamma}_{ij}\dot{\theta}_i\dot{\theta}_j=0 \qquad (k=1,\ldots,n)$$

gives the "straight-lines" $\theta=\theta(t)$ with respect to the α-connection. When $\alpha=0$ these "straight-lines" are also the information geodesic curves. This is not necessarily so when $\alpha\neq 0$ for, in this case, the α-connection is not compatible with the metric tensor g_{ij}.

The theory of α-connections and their curvatures seems to be particularly applicable in elucidating the structures of the exponential families as well as of the curved exponential families of distributions. An exponential family

may be written in the form (3.6)-(3.7) by choosing natural parameters $\theta=(\theta_1,\ldots,\theta_n)$ which are uniquely determined within affine transformations. In this case, $(T_1,\ldots,T_n)$ constitutes a sufficient statistic for the family and has a covariance matrix V which equals to I. In particular, the corresponding Cramer-Rao lower bound, in property 3^o(i) of the previous section, is always attained. Moreover, the natural parameter space Θ is convex, and, by (3.8), ψ is convex on Θ. A use of (3.5)-(3.10) shows that the α-Riemann-Christoffel curvature tensor of the space is given by

$$\overset{\alpha}{R}_{ijk\ell}=\frac{1-\alpha^2}{2}[T_{jrk}T_{im\ell}-T_{jr\ell}T_{imk}]g^{mr}.$$

Initially, this formula is valid only for the natural coordinate system. However, since the formula is given by means of a tensorial equation, its validity does not depend on a particular choice of the coordinates. It follows that for any exponential family P_Θ

$$\overset{\alpha}{R}_{ijk\ell}=(1-\alpha^2)R_{ijk\ell},$$

and hence the Efron and the Dawid connections (i.e. when $\alpha=1$ and $\alpha=-1$) render the space Θ as "flat" (or with an "absolute parallism").

The curved exponential families can be embedded in the exponential families as subspaces (Efron [13,14]). Using this observation, one shows that these families posses vari-

ous dualistic structures: The Barndorff-Nielsen duality [4] associated with the Legendre transformation, the α-$(-\alpha)$ duality [1] between two kinds of connections and the α-$(-\alpha)$ duality [1] between two kinds of curvatures. As shown by Amari [1], these dualities are intimately connected and, moreover, that the second-order information loss is expressed in terms of the curvatures of the statistical model and the estimator. We refer to Amari [1], Barndorff-Nielsen [4], Dawid [11], Efron [13,14] and Reeds [28] for a more detailed account on these statistical connections. For the general study of connections and curvatures, we refer to the books of Eisenhart [15,16], Hicks [18], Laugwitz [20] and Schouten [29].

4. Informative Geometry of Specific Families of Distributions.

An informative geometry of distributions $p(\cdot|\theta)\varepsilon P_\Theta$ is the geometry associated with the natural affine connection Γ_{ijk} of the information metric ds^2. We shall briefly describe the informative geometrics of certain well-known families of distribution P_θ. This description includes the evaluations of the curvature and the Rao distance for the family P_θ. Note, however, that, as mentioned in the previous section, the α-curvature based on the α-connection of Amari [1] is $(1-\alpha^2)$ times the present information curvature,

provided P_Θ is an exponential family.

4.1. Univariate Distributions.

Here Θ is an interval and $ds^2(\theta)=g(d\theta)^2$ with

$$g\equiv g(\theta)=g_{11}=-E_\theta(\partial^2 \log p) \qquad (p(\cdot|\theta)\in P_\Theta).$$

The curvature is always zero and the connection Γ_{111} is $g'(\theta)$. The latter can be made to vanish identically by reparametrizing $\theta\in\Theta$ to $s\in\Theta^*$, where $(s'(\theta))^2=g(\theta)$. The Rao distance of $a,b\in\Theta$ is given by

$$S(a,b)=\left|\int_a^b \sqrt{g(\theta)}\,d\theta\right|.$$

For example, for a one-parameter exponential family

$$p(x|\theta)=\exp\{T(x)+t(x)\phi(\theta)-\psi(\theta)\}$$

we find that $g=(\phi')^2\sigma^2>0$ where $\sigma^2=E_\theta[(T-\omega)^2]$ with $\omega=E_\theta(T)$, and, moreover

$$\omega=\psi'/\phi' \quad , \quad \sigma^2=\omega'/\phi'.$$

A special case is the (generalized) Weibull distribution

$$(4.1) \qquad p(x|\theta)=T'(x)\phi(\theta)\exp\{-T(x)\phi(\theta)\} \qquad (x>0,\ \theta\in\Theta)$$

with respect to the Lebesgue measure on $\chi=\mathbb{R}_+$. Here T is a non-negative differentiable function on χ with $T(0)=0$, and is monotonically increasing to ∞. We also assume that

$\phi(\theta)>0$ for every $\theta\varepsilon\Theta$. In this case $g=\{(\log\phi)'\}^2$ and thus

$$S(a,b)=\left|\log(\phi(a)/\phi(b))\right| \qquad (a,b\varepsilon\Theta).$$

We now list some other nondiscrete cases:

1° Gamma Distribution. Here

$$p(x|\theta)=\frac{1}{\Gamma(r)}x^{r-1}e^{-x\theta}\theta^{r}$$

with respect to the Lebesgue measure on $\chi=\mathbb{R}_+$. The parameter space Θ is $\mathbb{R}_+$ and $r>0$ is the index of the gamma distribution. In this case

$$S(a,b)=\sqrt{r}\left|\log(a/b)\right| \qquad (a,b\varepsilon\mathbb{R}_+).$$

2° Weibull Distribution.

$$p(x|\theta)=rx^{r-1}\theta e^{-x^{r}\theta}$$

with respect to the Lebesgue measure on $\chi=\mathbb{R}_+$. Here $\Theta=\mathbb{R}_+$ and $r>0$ is the index of the Weibull distribution. This is a special case of (4.1), and

$$S(a,b)=\left|\log(a/b)\right| \qquad (a,b\varepsilon\mathbb{R}_+).$$

3° Pareto Distribution.

$$p(x|\theta)=\theta r^{\theta}x^{-(\theta+1)} \qquad (x\geq r,\ \theta\varepsilon\mathbb{R}_+)$$

with respect to the Lebesgue measure on $\chi=[r,\infty)$, $r>0$. As before

$$S(a,b)=|\log(a/b)| \qquad (a,b\in\mathbb{R}_+)$$

4° Power Function Distribution.

$$p(x|\theta)=\theta r^{-\theta}x^{\theta-1} \qquad (0<x\leq r\ ,\ \theta\in\mathbb{R}_+)$$

with respect to the Lebesgue measure on $\chi=(0,r]$, $r>0$. Again,

$$S(a,b)=|\log(a/b)| \qquad (a,b\in\mathbb{R}_+).$$

5° Fixed-Mean Normal Distribution.

$$p(x|\theta)=N(x|r,\theta^2)=\frac{1}{(2\pi)^{1/2}\theta}\ e^{-(x-r)^2/2\theta^2} \qquad (x\in\mathbb{R})$$

with a fixed mean $r\in\mathbb{R}$ and variances θ^2, $\theta\in\Theta\equiv\mathbb{R}_+$. Then

$$S(a,b)=\sqrt{2}|\log(a/b)| \qquad (a,b\in\mathbb{R}_+).$$

6° Fixed-Variance Normal Distribution.

$$p(x|\theta)=N(x|\theta,r^2)=\frac{1}{(2\pi)^{1/2}r}e^{-(x-\theta)^2/2r^2} \qquad (x\in\mathbb{R})$$

with a fixed variance r^2, $r>0$ and means $\theta\in\Theta\equiv\mathbb{R}$. Then

$$S(a,b)=|a-b|/r \qquad (a,b\in\mathbb{R}).$$

We now list some discrete cases.

7° Poisson Distribution.

$$p(x|\theta)=e^{-\theta}\theta^x/x! \qquad (x\in\mathbb{Z}_+,\ \theta\in\mathbb{R}_+)$$

where $\mathbb{Z}_+=\{0,1,2,\ldots\}$. Then

$$S(a,b)=2\left|\sqrt{a}-\sqrt{b}\right| \qquad (a,b\in\mathbb{R}_+).$$

8° Negative Binomial Distribution.

$$p(x|\theta)=\frac{\Gamma(x+r)}{x!\Gamma(r)}\theta^x(1-\theta)^r \qquad (x\in\mathbb{Z}_+,\ 0<\theta<1)$$

with index $r>0$ and $\Theta=(0,1)$. The Rao distance is

$$S(a,b)=2\sqrt{r}\ \cosh^{-1}\left(\frac{1-\sqrt{ab}}{\sqrt{(1-a)(1-b)}}\right) \qquad (a,b\in\Theta).$$

Alternatively

$$S(a,b)=2\sqrt{r}\ \log\frac{1-\sqrt{ab}+\left|\sqrt{a}-\sqrt{b}\right|}{\sqrt{(1-a)(1-b)}} \qquad (a,b\in\Theta).$$

9° Binomial Distribution.

$$p(x|\theta)=\binom{N}{x}\theta^x(1-\theta)^{N-x} \qquad (x\in\{0,1,\ldots,N\},\ 0<\theta<1)$$

with $\Theta=(0,1)$ and $N\geq 1$ is an integer. In this case

$$S(a,b)=2\sqrt{N}\ \cos^{-1}\{\sqrt{ab}+\sqrt{(1-a)(1-b)}\}$$

or, equivalently,

$$S(a,b)=2\sqrt{N}\left|\sin^{-1}\sqrt{a}-\sin^{-1}\sqrt{b}\right| \qquad (a,b\in\Theta).$$

The distance without the factor $2\sqrt{N}$ is also called the "Hellinger-Bhattacharyya distance" (see [3,5,12]) for the binomial distribution.

4.2. Bivariate Distributions.

Here Θ, for $p(\cdot|\theta)\varepsilon P_\Theta$, is of dimension n=2. In this case, the first information curvature tensor $R_{ijk\ell}$ has only one independent component R_{1212}. The latter coincides with the Gaussian curvature κ. As an example, we describe the 2-dimenisonal geometry of the classical normal distribution $N(\mu,\sigma^2)$ with means μ and variances σ^2 ($\mu\varepsilon\mathbb{R}$, $\sigma\varepsilon\mathbb{R}_+$). Other examples are described in 4.3 below.

For the normal distribution

$$p(x|\theta)=N(x|\mu,\sigma^2)=\frac{1}{(2\pi)^{1/2}\sigma}e^{-(x-\mu)^2/2\sigma^2} \qquad (x\varepsilon\mathbb{R}),$$

we have $\Theta=\mathbb{R}\times\mathbb{R}_+$ and $\theta=(\mu,\sigma^2)\varepsilon\Theta$. The information metric is

$$ds^2=2\sigma^{-2}[(\frac{d\mu}{\sqrt{2}})^2+(d\sigma)^2].$$

and the curvature is $\kappa\equiv-2^{-1}$. Letting $\mu^*=\mu/\sqrt{2}$ and introducing the complex variable $z=\mu^*+i\sigma$ we find that Θ becomes the upper-half plane $\{z\varepsilon\mathbb{C}: \mathrm{Im}z>0\}$ and $ds^2=2\sigma^{-2}dzd\bar{z}$, effectively the Poincare metric. The geodesic curves are the "semi-circles"

$$z=a+re^{i\phi} \quad , \; r>0 \quad , \; 0<\phi<\pi,$$

where a is a real constant. This family includes the half-lines Rez=const., $z\varepsilon\Theta$, as limiting cases, corresponding to $r\to\infty$. The Rao distance S(1,2) between (μ_1,σ_1^2) and (μ_2,σ_2^2) of

θ, equivalently between $N(\mu_1,\sigma_1^2)$ and $N(\mu_2,\sigma_2^2)$, is

$$S(1,2)=\sqrt{2}\,\log\frac{1+\delta(1,2)}{1-\delta(1,2)}$$

where

$$\delta(1,2)=\left|\frac{(\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2}\right|^{1/2}.$$

Note that always $0\le\delta(1,2)<1$. An alternative expression for the distance is

$$S(1,2)=2\sqrt{2}\,\tanh^{-1}\delta(1,2).$$

We note that when $\mu_1=\mu_2$, the geodesic curve connecting (μ_1,σ_1^2) and (μ_1,σ_2^2) lies on the straight line μ=const., and the distance is $S(1,2)=\sqrt{2}\,|\log(\sigma_1/\sigma_2)|$ which is identical to that in 4.1(5^o). On the other hand, when $\sigma_1=\sigma_2$, the present distance does not agree with the distance in 4.1(6^o), since σ=const. is not a geodesic curve of the present metric (see also [1,3,7]).

4.3. Multivariate Distributions.

We first describe some discrete cases. To do so we employ the following multinomial notation: For $\theta=(\theta_1,\ldots,\theta_n)\in\mathbb{R}^n$ and $\alpha=(\alpha_1,\ldots,\alpha_n)\in\mathbb{Z}_+^n$, we let

$$|\theta|=|\theta_1|+\cdots+|\theta_n|\ ,\ \theta^\alpha=\theta_1^{\alpha_1}\cdots\theta_n^{\alpha_n}\ ,\ \alpha!=\alpha_1!\cdots\alpha_n!,$$

and hence $|\alpha|=\alpha_1+\cdots+\alpha_n$, and $|\theta|=\theta_1+\cdots+\theta_n$ if θ is also in

$\mathbb{R}^n_+$. If $y=(y_1,\ldots,y_n)$ is another vector in $\mathbb{R}^n$, then

$$\langle\theta,y\rangle=\theta_1y_1+\ldots+\theta_ny_n.$$

The vector $(1,\ldots,1)$ of $\mathbb{R}^n$ is denoted by $\mathbb{1}$.

At the present, the sample space χ is a subset of $\mathbb{Z}^n_+$ with a counting measure and the parameter space is of the form

$$\Theta_\rho=\{\theta\in\mathbb{R}^n_+:|\theta|<\rho\} \qquad (0<\rho\leq\infty).$$

A typical example is as follows: Let F be analytic with the power expansion

$$F(t)=\sum_{m=0}^{\infty} b_m t^m \qquad (-\rho<t<\rho), \tag{4.2}$$

such that $b_m>0$ for every $m\in Z_+$. Consider the probability distribution

$$p(\alpha|\theta)=\frac{|\alpha|!}{\alpha!}b_{|\alpha|}\frac{\theta^\alpha}{F(|\theta|)} \qquad (\alpha\in\mathbb{Z}^n_+,\ \theta\in\Theta_\rho). \tag{4.3}$$

The metric tensor is then

$$g_{ij}(\theta)=f(|\theta|)[\theta_i^{-1}\delta_{ij}+h(|\theta|)] \qquad (\theta\in\Theta_\rho),$$

where

$$f(t)=(\log F)'(t)\ ,\ h(t)=(\log f)'(t).$$

The first information curvature tensor in Θ_ρ is given by

$$R_{ijk\ell}(\theta)=-(4\theta_i\theta_j)^{-1}f(|\theta|)\{h(|\theta|)(\delta_{ik}\delta_{j\ell}-\delta_{i\ell}\delta_{jk})$$

$$+h'(|\theta|)\theta_i(\delta_{j\ell}-\delta_{jk})+h'(|\theta|)\theta_j(\delta_{ik}-\delta_{i\ell})\},$$

while the information curvature is

$$\kappa(\theta:x,y)=-\frac{1}{4[f(|\theta|)]^3}\{f(|\theta|)f'(|\theta|)$$

$$+H(|\theta|)[h(|\theta|)+\frac{A(x,y,y)}{<y,\theta>A(x,y,\underline{I})+<x,\theta>A(y,x,\underline{I})}]^{-1}\}$$

where

$$H(t)=f(t)f''(t)-2[f'(t)]^2$$

and

$$A(x,y,z)=<x,x><y,z>-<x,y><x,z> \qquad (x,y,z\varepsilon R^n).$$

This curvature is constant if and only if it is isotoropic, i.e. if and only if $H(t)\equiv 0$ for $-\rho<t<\rho$. The latter is equivalent to either $F(t)=ae^{rt}$ or $F(t)=a(b-t)^{-r}$ where a, b and r are positive constants with $\rho\leq b<\infty$. This gives, effectively, either the independent Poisson distribution with $\kappa\equiv 0$ or the negative multinomial distribution with $\kappa\equiv-(4r)^{-1}$ (see 1^o and 2^o below).

To find the geodesic curves for the distribution in (4.3), we introduce the additional functions,

$$L(t)=-\frac{1}{f(t)+tf'(t)}[f''(t)-2f'(t)h(t)]$$

and

$$M(t)=\frac{1}{f(t)+tf'(t)}[3f(t)f'(t)+tf(t)f''(t)+t(f'(t))^2].$$

The equations for the geodesic curves $\theta=\theta(s)$ are then

$$\frac{\ddot{\theta}_k}{\theta_k}-\frac{1}{2}\left(\frac{\dot{\theta}_k}{\theta_k}\right)^2+h(|\theta|)|\theta|^{\cdot}-\frac{1}{2}L(|\theta|)[|\theta|^{\cdot}]^2=0 \qquad (k=1,\ldots,n),$$

$$2f(|\theta|)|\theta|^{\cdot\cdot}+M(|\theta|)[|\theta|^{\cdot}]^2=1$$

and

$$f(|\theta|)\sum_{k=1}^{n}\frac{(\dot{\theta}_k)^2}{\theta_k}+f'(|\theta|)[|\theta|^{\cdot}]^2=1.$$

We now list some specific examples.

1°. Independent Poisson Distributions.

$$p(\alpha|\theta)=\frac{1}{\alpha!}e^{-|\theta|}\theta^{\alpha} \qquad (\alpha\in\mathbb{Z}_+^n,\ \theta\in\Theta_\infty).$$

This is a special case of (4.2) with $F(t)=e^t$. The metric tensor is

$$g_{ij}=\theta_i^{-1}\delta_{ij}$$

and the curvature is $\kappa\equiv 0$. The space Θ_∞ with respect to this metric is, therefore, essentially euclidean. The Rao distance for this distribution is

$$S(a,b)=2\{\sum_{k=1}^{n}(\sqrt{a_i}-\sqrt{b_i})^2\}^{1/2} \qquad (a,b\in\Theta_\infty),$$

effectively the "Hellinger distance" [7,8,12,23]. This example, together with 4.1(7°), provides also an illustration to property 1° of section 2.

2° Negative Multinomial Distributions.

$$p(\alpha|\theta)=\frac{\Gamma(|\alpha|+r)}{\alpha!\Gamma(r)}\theta^{\alpha}(1-|\theta|)^{r} \qquad (\alpha\varepsilon Z_{+}^{n}\ ,\ \theta\varepsilon\Theta_{1})$$

with index $r>0$. This is a special case of (4.2) with $F(t)=(1-t)^{-r}$. The metric tensor is

$$g_{ij}=\frac{r}{1-|\theta|}(\theta_{i}^{-1}\delta_{ij}+\frac{1}{1-|\theta|}) \qquad (\theta\varepsilon\Theta_{1}),$$

and

$$R_{ijk\ell}=-\frac{r}{4(1-|\theta|)^{2}\theta_{i}\theta_{j}}\{\delta_{ik}\delta_{j\ell}-\delta_{i\ell}\delta_{jk}+\frac{1}{1-|\theta|}[\theta_{i}(\delta_{j\ell}-\delta_{jk})+\theta_{j}(\delta_{ik}-\delta_{i\ell})]\}$$

with

$$\kappa\equiv-\frac{1}{4r}.$$

It follows that Θ_{1} with the above metric is locally isometric to the "Poincaré hyperbloic space". In particular, for any two points $a,b\varepsilon\Theta_{1}$ there exists a unique geodesic curve in Θ_{1}, with respect to the metric, connecting the points (see, for example, Hicks [18]).

The geodesic curves $\theta=\theta(s)$ are given by the "hyperbolas"

$$\theta_{k}=\{A_{k}\tanh\frac{1}{2\sqrt{r}}(s+B_{n+1})+B_{k}\}^{2} \qquad (k=1,\ldots,n)$$

where A_{j}, $B_{j}(j=1,\ldots,n)$ and B_{n+1} are constants satisfying

$$\sum_{k=1}^{n}A_{k}B_{k}=0 \quad , \quad \sum_{k=1}^{n}(A_{k}^{2}+B_{k}^{2})=1$$

and

$$0<\left|A_k \tanh\frac{B_{n+1}}{2\sqrt{r}}+B_k\right|<1 \qquad (k=1,\ldots,n).$$

This family includes the lines $\theta_k=B_k^2$ $(k=1,\ldots,n)$ as limiting cases. The Rao distance is

$$S(a,b)=2\sqrt{r}\,\cosh^{-1}\left|\frac{1-\sum_{k=1}^{n}\sqrt{a_k b_k}}{\sqrt{(1-|a|)(1-|b|)}}\right| \qquad (a,b\varepsilon\Theta_1).$$

An alternative expression may be obtained by using the identity $\cosh^{-1}x=\log[x+\sqrt{x^2-1}]$, $x\geq 1$. The expressions agree with those in 4.1(8^o) when n=1 (see also Oller and Cuadras [22]).

3^o. Multinomial Distributions.

$$p(\alpha|\theta)=\frac{N!}{(N-|\alpha|)!}\,\frac{\theta^\alpha}{\alpha!}(1-|\theta|)^{N-|\alpha|} \qquad (\alpha\varepsilon Z_+^n\ ,\ |\alpha|<N\ ;\ \theta\varepsilon\Theta_1)$$

with an integer $N\geq 1$ and sample space $\chi=\{\alpha\varepsilon Z_+^n:\ |\alpha|\leq N\}$. The metric tensor is

$$g_{ij}=N\left(\theta_i^{-1}\delta_{ij}+\frac{1}{1-|\theta|}\right) \qquad (\theta\varepsilon\Theta_1),$$

and

$$R_{ijk\ell}=\frac{N}{4\theta_i\theta_j}\left\{\delta_{ik}\delta_{j\ell}-\delta_{i\ell}\delta_{jk}+\frac{1}{1-|\theta|}\left[\theta_i(\delta_{j\ell}-\delta_{jk})+\theta_j(\delta_{ik}-\delta_{i\ell})\right]\right\}$$

with

$$\kappa\equiv\frac{1}{4N}.$$

This gives, effectively, the spherical geometry. In fact,

upon putting $\vartheta_{n+1}=1-|\theta|$ and introducing the new variables $y_i=\theta_i^{1/2}$ $(i=1,\ldots,n+1)$, we find that the metric becomes

$$ds^2(y)=4N\sum_{i=1}^{n+1}(dy_i)^2$$

and Ω_1 is mapped onto the positive portion of the $(n+1)$-dimensional unit sphere $Y=\{y=(y_1,\ldots,y_{n+1})\in R_+^{n+1}: y_1^2+\cdots+y_{n+1}^2=1\}$. It follows from the spherical representation that the geodesic curves $\theta=\theta(s)$ are the "great circles"

$$\theta_k(s)=(A_k\cos\frac{s}{2\sqrt{N}}+B_k\sin\frac{s}{2\sqrt{N}})^2 \qquad (k=1,\ldots,n+1)$$

with

$$\theta_{n+1}=1-(\theta_1+\cdots+\theta_n)$$

and with constants $(A_1,\ldots,A_{n+1})$, $(B_1,\ldots,B_{n+1})$ satisfying

$$\sum_{k=1}^{n+1}A_k^2=\sum_{k=1}^{n+1}B_k^2=1\ ,\ \sum_{k=1}^{n+1}A_kB_k=0.$$

The Rao distance is

$$S(a,b)=2\sqrt{N}\ \cos^{-1}(\sum_{k=1}^{n+1}\sqrt{a_kb_k}) \qquad (a,b\in\Theta_1)$$

with

$$a_{n+1}=1-|a|\ ,\ b_{n+1}=1-|b|,$$

effectively the "Hellinger-Bhattacharyya distance" [5,6] (see also [3,24]). This agrees with 4.1(9^o) when n=1.

The nondiscrete cases that we shall describe here are

those associated with the normal distribution

$$N(x:\mu,\Sigma)=\frac{1}{(2\pi)^{n/2}}\frac{1}{|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\} \qquad (x\in\mathbb{R}^n),$$

with mean (column) vector μ and a variance-covariance matrix Σ. We shall use the standard matrix notation: $M(n,\mathbb{R})$ is the space of all $n\times n$ real matrices, $S(n,\mathbb{R})=\{A\in M(n,\mathbb{R}): A'=A\}$ the subspace of symmetric matrices, $GL(n,\mathbb{R})$ the group of all non-singular matrices in $M(n,\mathbb{R})$, $P(n,\mathbb{R})$ the subset of all positive-definite symmetric matrices in $GL(n,\mathbb{R})$. The inner product and norm on $M(n,\mathbb{R})$ are given by

$$\langle A,B\rangle=\mathrm{tr}(AB') \ , \ ||A||=\{\langle A,A\rangle\}^{1/2} \qquad (A,B\in M(n,\mathbb{R})),$$

and

$$[A,B]=AB-BA \qquad (A,B\in M(n,\mathbb{R}))$$

stands for the commutator of A and B.

For the normal distribution $p(\cdot|\theta)=N(\cdot|\mu,\Sigma)$, Θ is $\mathbb{R}^n\times P(n,\mathbb{R})$ and $\theta=(\mu,\Sigma)\in\Theta$. The information metric is then

$$(4.4) \quad ds^2=(d\mu)'\Sigma^{-1}(d\mu)+\frac{1}{2}\mathrm{tr}\{(\Sigma^{-1}d\Sigma)^2\} \qquad ((\mu,\Sigma)\in\Theta).$$

We note that

$$(4.5) \quad \mathrm{tr}\{(\Sigma^{-1}d\Sigma)^2\}=||\Sigma^{-1/2}d\Sigma\Sigma^{-1/2}||^2 \qquad (\Sigma\in P(n,\mathbb{R})).$$

The geodesic curves are determined by

$$(4.6)\qquad \begin{cases} \Sigma^{-1}\dot{\mu}=c & (c\varepsilon R^n), \\ (\Sigma^{-1}\dot{\Sigma})^{\cdot}+cc'\Sigma=0, \\ c'\Sigma c+\frac{1}{2}tr\{(\Sigma^{-1}\dot{\Sigma})^2\}=1, \end{cases}$$

where c is a constant vector in R^n. We also note that for any (a,A) in $R^n\times GL(n,R)$, the mapping $(\mu,\Sigma)\to(A'\mu+a,\ A\ \Sigma A)$ establishes a homeomorphism of Θ onto Θ which is also an isometry with respect to the information metric ds^2. Consequently, the Rao distance between (μ_1,Σ_1) and (μ_2,Σ_2) in Θ satisfies

$$S(\mu_1,\Sigma_1:\mu_2,\Sigma_2)=S(A'\mu_1+a,A'\Sigma_1A:A'\mu_2+a,A'\Sigma_2A)$$

for any $(a,A)\varepsilon R^n\times GL(n,R)$. In particular, the above Rao distance $S(1,2)$ admits the form

$$S(1,2)=S(0,I:\Sigma_1^{-1/2}(\mu_2-\mu_1),\ \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}).$$

Explicit expressions for the geodesic curves and the Rao distance in this general setting are not available. We therefore only describe some special cases.

4°. Fixed-Variance-Covariance Normal Distributions.

In this case we consider the family of normal distributions $N(\cdot|\mu,\Sigma_o)$ with a fixed variance-covariance matrix Σ_o. In this case Θ is R^n and the information metric is

$$ds^2=(d\mu)'\Sigma_o^{-1}(d\mu)\qquad(\mu\varepsilon R^n),$$

which is essentially the euclidean metric on $\mathbb{R}^n$, since $\Sigma_o \in P(n,\mathbb{R})$ is constant. The Rao distance is therefore

$$S(\mu_1,\mu_2)=\{(\mu_1-\mu_2)' \Sigma_o^{-1}(\mu_1-\mu_2)\}^{1/2} \qquad (\mu_1,\mu_2 \in \mathbb{R}^n).$$

This is the familiar "Mahalanobis distance" [18] and it agrees with the distance in 4.1(6^o) when n=1. Note, however, that, as is mentioned also in 4.2, the present distance cannot be regarded as the restriction of the Rao distance for the entire manifold $\mathbb{R}^n \times P(n,\mathbb{R})$ to $\mathbb{R}^n \times \{\Sigma_o\}$. This is because the curve (μ,Σ_o), as (4.6) shows, is not a geodesic curve of the metric in (4.4).

5°. Fixed-Mean Vector Normal Distributions.

Here we consider the family of normal distributions $N(\cdot|\mu_o,\Sigma)$ with a fixed mean vector μ_o. In this case Θ is $P(n,\mathbb{R})$ and the information metric is

$$(4.7) \qquad ds^2 = \frac{1}{2}tr\{(\Sigma^{-1}d\Sigma)^2\} \qquad (\Sigma \in P(n,\mathbb{R})).$$

Moreover, the geodesic curves $\Sigma=\Sigma(s)$ of this metric may be determined from (4.6) with c=0. This gives $(\Sigma^{-1}\dot{\Sigma})^{\cdot}=0$ and the normalization $tr\{(\Sigma^{-1}\dot{\Sigma})^2\}=2$. The solution $\Sigma(s)$ must belong to $P(n,\mathbb{R})$. Consequently the most general geodesic curve is of the form

$$(4.8) \qquad \Sigma(s)=A' e^{sB} A$$

where A and B are constant matrices such that

(4.9) $\quad A\varepsilon GL(n,\mathbb{R})\ ,\ B\varepsilon S(n,\mathbb{R})\ ,\ ||B||^2=tr(B^2)=2.$

The group of automorphisms G of $P(n,\mathbb{R})$ onto itself is generated by $\Sigma\mapsto\Sigma^{-1}$ and $\Sigma\mapsto A'\Sigma A$ where $A\varepsilon GL(n,\mathbb{R})$. This group is "transitive", i.e. for any $\Sigma_1,\Sigma_2\varepsilon P(n,\mathbb{R})$ there exists an $f\varepsilon G$ such that $f(\Sigma_1)=\Sigma_2$ (just choose $A=\Sigma_1^{-1/2}\Sigma_2^{1/2}\varepsilon GL(n,\mathbb{R})$). Moreover, for a given $\Sigma_1\varepsilon P(n,\mathbb{R})$, the automorphism $f\varepsilon G$ given by $f(\Sigma)=\Sigma_1\Sigma^{-1}\Sigma_1$, $\Sigma\varepsilon P(n,\mathbb{R})$, satisfies $f=f^{-1}$ and $f(\Sigma_1)=\Sigma_1$. Consequently, the parameter space $\Theta=P(n,\mathbb{R})$ is a "symmetric space". It is also easily verified that the group G coincides with the group of isometries of the metric ds^2 on $P(n,\mathbb{R})$. The group G forms a subgroup of the "Siegel sympletic group" [30] which acts on the "Siegel upper-half space" $S(n,\mathbb{R})+iP(n,\mathbb{R})$. This gives an alternative description for the geodesic curves which is equivalent to that of (4.8)-(4.9), namely that the most general geodesic curve $\Sigma=\Sigma(s)$ is of the form

$$f(\Sigma(s))=\mathrm{diag}[e^{\nu_1 s},\ldots,e^{\nu_n s}] \qquad (f\varepsilon G)$$

where f is an arbitrary automorphism of $P(n,\mathbb{R})$ and $\nu_1,\ldots,\nu_n$ are arbitrary non-negative numbers with $\nu_1^2+\cdots+\nu_n^2=2$.

The Rao distance $S(1,2)$ between Σ_1 and Σ_2 of $P(n,\mathbb{R})$ is easily determined from the above geodesic equations. The geodesic curve $\Sigma=\Sigma(s)$ along these two points, with $\Sigma(0)=\Sigma_1$, $\Sigma(s_o)=\Sigma_2$ and $S_o\equiv S(1,2)$, satisfies (4.8)-(4.9) in the inter-

val $0\leq s\leq s_o$ with

$$A'A=\Sigma_1 \ , \ e^{s_o B}=(A^{-1})'\Sigma_2 A^{-1}.$$

Consequently,

$$(4.10)\quad S(1,2)=\frac{1}{\sqrt{2}}||\log\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}|| \qquad (\Sigma_1,\Sigma_2\epsilon P(n,\mathbb{R})).$$

We note that

$$(4.11)\quad ||\log\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}||^2=tr\{\log^2\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\}=\sum_{k=1}^{n}\log^2\lambda_k,$$

where

$$(4.12)\qquad \lambda_k=\lambda_k(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}) \qquad (k=1,\ldots,n)$$

are the positive eigenvalues of the positive-definite matrix $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$ (note also the symmetry between Σ_1 and Σ_2 in (4.11)). Equivalently, $\lambda_1,\ldots,\lambda_n$ are the singular-values of $\Sigma_2\Sigma_1^{-1}$, or of $\Sigma_1^{-1}\Sigma_2$, and they are determined uniquely as the solutions $\lambda=\lambda_k(k=1,\ldots,n)$ of the determinantal equation

$$|\lambda\Sigma_2-\Sigma_1|=0.$$

Other alternative expressions for the Rao distance S(1,2) in (4.10) are also available. For this purpose we introduce the symmetric matrix

$$T_{12}=(\Sigma_1-\Sigma_2)(\Sigma_1+\Sigma_2)^{-1}$$

and define $R=R(\Sigma_1,\Sigma_2)$ by

$$R=T_{12}^2 \ (\equiv T_{12}'T_{12}\equiv T_{12}T_{12}').$$

Then R is symmetric and semi positive-definite; it is positive-definite if $\Sigma_1\neq\Sigma_2$ and is zero otherwise. Consequently, the eigenvalues $r_k=r_k(R)$ of R are related to the eigenvalues λ_k in (4.12) by

$$r_k=\left(\frac{1-\lambda_k}{1+\lambda_k}\right)^2 \ , \ 0\leq r_k<1 \qquad (k=1,\ldots,n).$$

In particular, the matrices $I+R^{1/2}$ and $I-R^{1/2}$ are members of $P(n,\mathbb{R})$. Moreover, since

$$\log^2\frac{1+r^{1/2}}{1-r^{1/2}}=4[\tanh^{-1}r^{1/2}]^2=4r\left(\sum_{k=0}^{\infty}\frac{r^k}{2k+1}\right)^2 \qquad (0\leq r<1)$$

and

$$\mathrm{tr}(R^j)=\sum_{k=1}^{n} r_k^j \qquad (j=0,1,\ldots),$$

we have, noting that $I+R^{1/2}$ and $I-R^{1/2}$ commute,

$$\log^2\frac{I+R^{1/2}}{I-R^{1/2}}=4[\tanh^{-1}R^{1/2}]^2=4R\left(\sum_{k=0}^{\infty}\frac{R^k}{2k+1}\right)^2,$$

and therefore

$$(4.13)\quad S^2(1,2)=\frac{1}{2}\mathrm{tr}\{\log^2\frac{I+R^{1/2}}{I-R^{1/2}}\}=\frac{1}{2}\sum_{k=1}^{n}\log^2\frac{1+r_k^{1/2}}{1-r_k^{1/2}}\ .$$

The components of the first curvature tensor at any $\Sigma\in P(n,\mathbb{R})$ are zero and thus the space $P(n,\mathbb{R})$ is essentially euclidean. The Rao distance in (4.10) or in (4.13) reduces to that in 4.1(5^o) when n=1 (see also [3]).

6^o. Independent Normal Distributions.

We consider a family of normal distributions

$N(\cdot|\mu,\Sigma:\Sigma_o)$ with varying mean vectors $\mu\epsilon\mathbb{R}^n$ and variance-covariance matrices $\Sigma\epsilon P(n,\mathbb{R})$ that commute with a fixed variance-covariance matrix $\Sigma_o\epsilon P(n,\mathbb{R})$, $\Sigma_o\neq I$. The parameter space Θ in this case is $\mathbb{R}^n\times\tilde{P}(n,\mathbb{R}:\Sigma_o)$, where

$$\tilde{P}(n,\mathbb{R}:\Sigma_o)=\{\Sigma\epsilon P(n,\mathbb{R}):\Sigma\Sigma_o=\Sigma_o\Sigma\}.$$

The last set contains the matrices I and Σ_o, and it contains all the powers Σ^m, $m=\pm1,\pm2,\ldots$, provided $\Sigma\epsilon\tilde{P}(n,\mathbb{R}:\Sigma_o)$. In addition, if Σ_1 and Σ_2 are any members of $\tilde{P}(n,\mathbb{R}:\Sigma_o)$ then so is $\Sigma_1\Sigma_2\Sigma_1$.

The fixed matrix Σ_o admits the decomposition

$$\Sigma_o=U_o'\Lambda_o^2U_o$$

where U_o is an orthogonal matrix,

$$U_o'U_o=U_oU_o'=I,$$

with positive elements in the diagonal, and Λ_o is a diagonal matrix

$$\Lambda_o=\text{diag}[\sigma_{o1},\ldots,\sigma_{on}] \qquad (\sigma_{ok}>0\ ;\ k=1,\ldots,n),$$

with positive elements. It follows from the well-known result on the simultaneous diagonalization of commuting symmetric matrices (see, for example, [27, p. 41]) that $\tilde{P}(n,\mathbb{R}:\Sigma_o)\equiv P(n,\mathbb{R}:U_o)$ where

$$P(n,\mathbb{R}:U_o)=\{\Sigma\epsilon P(n,\mathbb{R}):\Sigma=U_o\Lambda^2U_o\}$$

with

$$\Lambda=\text{diag}[\sigma_1,\ldots,\sigma_n] \ (\sigma_k>0 \ , \ k=1,\ldots,n).$$

This shows that $\tilde{P}(n,\mathbb{R}:\Lambda_o^2)\equiv P(n,\mathbb{R}:I)$ is the set of all diagonal matrices in $P(n,\mathbb{R})$ and that it is isomorphic to $\tilde{P}(n,\mathbb{R}:\Sigma_o)$. In particular, $\tilde{P}(n,\mathbb{R}:\Sigma_o)$ is an n-dimensional submanifold of the full $n(n+1)/2$-dimensional manifold $P(n,\mathbb{R})$.

The mapping $x\mapsto U_o x$ constitutes an isometry of the sample space $\chi=\mathbb{R}^n$ onto itself and preserves the Lebesgue measure on $\mathbb{R}^n$. Therefore, the given family of distributions $N(\cdot|\mu,\Sigma:\Sigma_o)$ is identical with

$$N(\cdot|\mu,\Lambda^2:U_o)\equiv N(\cdot|\nu_1,\sigma_1^2)\cdots N(\cdot|\nu_n,\sigma_n^2) \tag{4.14}$$

where

$$\nu=(\nu_1,\ldots,\nu_n)'=U_o\mu, \tag{4.15}$$

$$\Lambda=\text{diag}[\sigma_1,\ldots,\sigma_n] \quad (\sigma_k>0:k=1,\ldots,n).$$

The given distributions are therefore products of n independent univariate normal distributions, and hence, by virtue of property 1^o of section 2, the analysis can be reduced to that found in 4.2. The parameter space Θ is now $(R\times\mathbb{R}_+)^n$ and $\theta=[(\nu_1,\sigma_1^2),\ldots,(\nu_n,\sigma_n^2)]'\in(\mathbb{R}\times\mathbb{R}_+)^n$, and the information metric is

$$ds^2=2\sum_{k=1}^{n}\sigma_k^{-2}[(\frac{d\nu_k}{\sqrt{2}})^2+(d\sigma_k)^2].$$

Letting $\nu_k^*=\nu_k/\sqrt{2}$ and introducing the complex variables $z_k=\nu_k^*+i\sigma_k$ $(k=1,\ldots,n)$ we find that Θ becomes the poly-upper-half plane $U^n=\{z=(z_1,\ldots,z_n)\in\mathbb{C}^n: \mathrm{Im}z_k>0,\ k=1,\ldots,n\}$ and

$$ds^2=2\sum_{k=1}^{n}\sigma_k^{-2}dz_k d\bar{z}_k \qquad (z\in U^n),$$

effectively the "Poincaré metric" of U^n.

The above metric is "hermitian" (see [19]), i.e. it is of the form

$$ds^2=g_{k\bar{j}}dz_k d\bar{z}_j,$$

where the summation convention has been used. Here $[g_{k\bar{j}}]$ is an $n\times n$ hermitian (i.e. $\bar{g}_{k\bar{j}}=g_{j\bar{k}}$; $k,j=1,\ldots,n$) matrix, defined on a complex manifold M of a complex dimension n. For a local coordinate system $z_1,\ldots,z_n$ of M with $z_k=x_k+iy_k$ $(k=1,\ldots,n)$, we introduce the complex-derivatives

$$\partial_k=(\partial_{x_k}-i\partial_{y_k})/2\ ,\ \bar{\partial}_k=(\partial_{x_k}+i\partial_{y_k})/2 \qquad (k=1,\ldots,n)$$

and the components of the "Ricci curvature" tensor

$$R_{k\bar{j}}=-2\partial_k\bar{\partial}_j\log G,$$

where G is the determinant of $[g_{k\bar{j}}]$. The components of the Riemann curvature tensor are now given by

$$R_{ij\bar{k}\bar{\ell}}=-\partial_k\bar{\partial}_\ell g_{i\bar{j}}+g^{r\bar{m}}\partial_k g_{i\bar{m}}\bar{\partial}_\ell g_{r\bar{j}},$$

while the mean Gaussian curvature is replaced by the "holo-

morphic sectional curvature"

$$\kappa(z:v)=\frac{2R_{i\bar{j}k\bar{\ell}}v_i\bar{v}_j v_k\bar{v}_\ell}{[g_{i\bar{j}}v_i\bar{v}_j]^2} \qquad (z\varepsilon M),$$

at $z\varepsilon M$ in the direction of $v=(v_1,\ldots,v_n)\varepsilon\mathbb{C}^n$. Here $[g^{r\bar{m}}]$ is the matrix-inverse of $[g_{r\bar{m}}]$.

For the metric under consideration we have

$$g_{k\bar{j}}=2\sigma_k^{-2}\delta_{kj}\ ,\ g^{k\bar{j}}=\frac{1}{2}\sigma_k^2\delta_{kj}$$

and

$$G=2^n[\sigma_1\cdots\sigma_n]^{-2}.$$

Therefore

$$R_{k\bar{j}}=-\frac{1}{2}g_{k\bar{j}}=-\sigma_k^{-2}\delta_{kj}.$$

Moreover,

$$R_{i\bar{j}k\bar{\ell}}=-\sigma_k^{-4}\delta_{ijk\ell}$$

where $\delta_{ijk\ell}$ is the tensor whose components are of value 1 if $i=j=k=\ell$ and 0 otherwise. It follows that the information holomorphic sectional curvature at $z\equiv[(\nu_1,\sigma_1^2),\ldots,(\nu_n,\sigma_n^2)]'$ in the direction of $v=(v_1,\ldots,v_n)\varepsilon\mathbb{C}^n$ is

$$\kappa(z:v)=-\frac{1}{2}\frac{\sum_{k=1}^{n}\sigma_k^{-4}|v_k|^4}{[\sum_{k=1}^{n}\sigma_k^{-2}|v_k|^2]^2}.$$

This curvature is independent of the mean vector $\nu=U_o\mu$, and it varies between $-1/2$ and $-1/2n$, a result consistent with 4.2.

The geodesic curves are the product of the "semi-circles"

$$\nu_k=\sqrt{2}(a_k+r_k\cos\phi_k) \ , \ \sigma_k=r_k\sin\phi_k$$

where

$$r_k>0 \ , \ 0<\phi_k<\pi \qquad (k=1,\ldots,n)$$

and $a_1,\ldots,a_n$ are real constants. This family includes products containing the half-lines $\nu_k=$cost. as limiting cases. The Rao distance $S(1,2)$ between $(\underline{\nu}_1,\underline{\sigma}_1^2)$ $=[(\nu_{11},\sigma_{11}^2),\ldots,(\nu_{1n},\sigma_{1n}^2)]'$ and $(\underline{\nu}_2,\underline{\sigma}_2^2)$ $=[(\nu_{21},\sigma_{21}^2),\ldots,(\nu_{2n},\sigma_{2n}^2)]'$, with the identification of of (4.14)-(4.15), is given by

$$S(1,2)=\sqrt{2}\left[\sum_{k=1}^{n}\log^2\frac{1+\delta_k(1,2)}{1-\delta_k(1,2)}\right]^{1/2}$$

where

$$\delta_k(1,2)=\left[\frac{(\nu_{1k}-\nu_{2k})^2+2(\sigma_{1k}-\sigma_{2k})^2}{(\nu_{1k}-\nu_{2k})^2+2(\sigma_{1k}+\sigma_{2k})^2}\right]^{1/2} \qquad (k=1,\ldots,n).$$

This distance reduces to that in 4.2 when n=1 (see also [7] for additional details).

5. Hilbert Space Embedding.

The intrinsic geometry of a space of distributions may be represented by means of an embedding in a Hilbert space. In order to describe this abstract approach we shall introduce some further notation.

For $\alpha \in \mathbb{R}$ we let

$$L^\alpha = \{p \in M: \int_X |p|^\alpha \, d\mu < \infty\} \quad (\alpha \neq 0),$$

$$L^0 = \{p \in M: \int_X (\log|p|)^2 \, d\mu < \infty\},$$

and for $0 < r < \infty$ we also consider the subsets

$$L^\alpha(r) = \{p \in L^\alpha: \int_X |p|^\alpha \, d\mu = r^\alpha\} \quad (\alpha \neq 0),$$

$$L^0(r) = \{p \in L^0: \int_X (\log |p|)^2 \, d\mu = r\}.$$

In this notation, L^2 is a Hilbert space with the inner product and norm

$$(p,q) = \int_X pq d\mu, \quad ||p|| = \sqrt{(p,p)} \quad (p,q \in L^2),$$

and $L^2(r)$ is the sphere of radius r in L^2. We also define

$$L_+^\alpha = L^\alpha \cap M_+, \quad P^\alpha(r) = L^\alpha(r) \cap M_+$$

and we write P^α for $P^\alpha(1)$. Thus, in the notation of Section 1, $L_+ = L_+^1$ and $P = P^1$.

For $\alpha \in \mathbb{R}$ and $p \in M_+$ we define

$$T_\alpha(p) = \frac{2}{|\alpha|} p^{\alpha/2} \quad (\alpha \neq 0), \quad T_0(p) = \log p \quad (p \in M_+). \tag{5.1}$$

Then T_α is a bijection of M_+ onto M_+ for any $\alpha \neq 0$ while T_0 is a bijection of M_+ onto M. Moreover, T_α embeds L_+^α into L^2 with

$$(5.2) \qquad T_\alpha(L_+) = L_+^2 \ , \ T_\alpha(P^\alpha) = P^2\left(\frac{2}{|\alpha|}\right) \quad (\alpha \neq 0)$$

and

$$(5.3) \qquad T_0(L_+^0) = L^2 \ , \ T_0(P^0) = P^2(1).$$

The induced distance on L_+^α is

$$(5.4) \qquad \rho_\alpha(p_1,p_2) = \frac{2}{|\alpha|}\,||p_1^{\alpha/2} - p_2^{\alpha/2}|| \quad (p_1,p_2 \in L_+,\ \alpha \neq 0)$$

and

$$(5.5) \qquad \rho_0(p_1,p_2) = ||\log p_1 - \log p_2|| \quad (p_1,p_2 \in L_+^0).$$

Here, the distance $\rho_1(p_1,p_2) = 2||\sqrt{p_1} - \sqrt{p_2}||$ is known as the "Hellinger distance" on $L_+ = L_+^1$. We also note that under some regularity conditions on $p_1,p_2 \in L_+^0$ we have

$$\rho_0(p_1,p_2) = \lim_{\alpha \to 0} \rho_\alpha(p_1,p_2).$$

Let $\theta = (\theta_1,\theta_2,\ldots)'$ be a set of real continuous parameters belonging to a parameter space Θ, a manifold embedded in some $\mathbb{R}^n$, $1 \leq n \leq \infty$. Here $\mathbb{R}^\infty = \ell^2 = \{(a_1,a_2,\ldots)' : \sum_{k=1}^{\infty} a_k^2 < \infty,$ $a_k \in \mathbb{R},\ k = 1,2,\ldots\}$. Let $F_\Theta^\alpha = \{p(.|\theta) \in L_+^\alpha : \theta \in \Theta\}$, $\alpha \in \mathbb{R}$, be a parametric family of positive distributions $p = p(.|\theta), \theta \in \Theta$, having suitable regularity properties. For example, we assume that

$$\partial_i p = \partial p(.|\theta)/\partial\theta_i \qquad (p = p(.|\theta) \in F_\Theta^\alpha)$$

is in M for every $\theta \in \Theta$ and each $i = 1,2,\ldots$. We also consider the subfamily $P_\Theta^\alpha = F_\Theta^\alpha \cap P^\alpha$ of F_Θ^α.

On F_θ^α we have

$$\rho_\alpha^2(p(.|\theta),\ p(.|\theta+\delta\theta)) = ds_\alpha^2(\theta),$$

to the second order infinitesimal displacements. Here ds_α^2 is the α-order information metric

$$ds_\alpha^2(\theta) = \int_X p^\alpha\ (d\log p)^2\ d\mu \qquad (\alpha \in \mathbb{R}),$$

where the dependence on $x \in X$ and $\theta \in \Theta$ in the integral has been supressed. This may also be written as

$$ds_\alpha^2(\theta) = d\theta' I_\alpha(\theta) d\theta$$

where

$$I_\alpha(\theta) = [g_{ij}^{(\alpha)}(\theta)]$$

is the α-order information matrix with

$$g_{ij}^{(\alpha)}(\theta) = \int_X p^\alpha\ (\partial_i \log p)(\partial_j \log p)\ d\mu.$$

This matrix is always semi positive-definite. It is positive-definite at $\theta \in \Theta$ if and only if the functions $\{\partial_i p\}$ are linearly independent over X. Note that $ds^2(\theta) = ds_1^2(\theta)$ and $I(\theta) = I_1(\theta)$ are the ordinary information metric and information matrix, respectively.

The geometries of L_+^α and P^α $(\alpha \in \mathbb{R})$ under the α-order metric ds_α^2 may be read off from the embedding T_α of L_+^α into L^2

$$q = T_\alpha(p) \qquad (p \in L_+^\alpha). \tag{5.6}$$

We then have

$$ds_\alpha^2(p) = ds_2^2(q) = \|dq\|^2 \quad (q \in L^2). \tag{5.7}$$

Here the parameter space Θ may be taken as a subset of L^2. The coordinates of a point q in L^2 may be determined by any

orthonormal basis $(e_1,e_2,\ldots)$ of L^2 via the Fourier-coefficients

$$q_k = (q,e_k) \qquad (k = 1,2,\ldots). \tag{5.8}$$

In this way the point $q \in L^2$ is identified with the point $(q_1,q_2,\ldots)'$ of ℓ^2 and we have

$$\|q\| = \sum_{k=1}^{\infty} q_k^2 \qquad (q \in L^2). \tag{5.9}$$

When Θ is L^2, the geometry under ds_2^2 is the usual euclidean geometry. The Riemann-Christoffel tensor of the first kind in identically zero and the geodesic curves $q[s] = q(\cdot|s) \in L^2$ are the "straight lines"

$$q[s] = as + b \qquad (q[s] \in L^2,\ s \in \mathbb{R})$$

where a and b are parameter-independent functions in L^2. The geodesic distance is then

$$S_2(q_1,q_2) = \rho_2(q_1,q_2) = \|q_1 - q_2\| \quad (q_1,q_2 \in L^2).$$

When, on the other hand, Θ is $L^2(r)$ $(0 < r < \infty)$, the geometry under ds_2^2 is the spherical geometry. In this case, the Riemann-Christoffel tensor of the first kind is given by

$$R_2(x,y:\ u,v) = (1/4)\{(x,u)(y,v) - (x,v)(y,u)\} \tag{5.10}$$

where $x,y,u,v \in L^2$. The mean Gaussian curvature is then

$$\kappa_2(x,y) = \frac{R_2(x,y:x,y)}{\|x\|^2\,\|y\|^2 - [(x,y)]^2} \equiv 1/4 \quad (x,y \in L^2).$$

To find the geodesic curves $q = q[s]$ $(0 \le s \le L)$ of this spherical geometry, we determine the solutions of the first variation equation

$$\delta \int_0^L \|\dot{q}[s]\| \, ds = 0.$$

subject to the constraint

$$\|q[s]\| = r \qquad (0 \leq s \leq L) \tag{5.12}$$

Here s is the arc-length parameter and thus we also have the normalization

$$\|\dot{q}[s]\| = 1 \qquad (0 \leq s \leq L). \tag{5.13}$$

For this purpose, we consider the Lagrangian

$$G(q,\dot{q}) = \|\dot{q}[s]\| + \lambda(s) \, [\, \|q[s]\| - r^2]$$

with the Lagrange multiplier $\lambda(s)$. Using an orthonormal basis $(e_1, e_2, \ldots)$ of L^2, the Lagrangian may be represented with the aid of (5.8)-(5.9) as

$$G(q,\dot{q}) = \{ \sum_{k=1}^{\infty} \dot{q}_k^2 \}^{\frac{1}{2}} + \lambda(s) \, \{ \sum_{k=1}^{\infty} q_k^2 - r^2 \}, \tag{5.14}$$

where $(q_1, q_2, \ldots)' \in \ell^2$ is the coordinatization of $q \in L^2$.

Thus we seek the extremum of

$$\int_0^L G(q,\dot{q}) \, ds$$

subject to the constraint (5.12) and the normalization (5.13) which, in view of (5.8)-(5.9), may be written as

$$\sum_{k=1}^{\infty} q_k^2 = r^2, \qquad \sum_{k=1}^{\infty} \dot{q}_k^2 = 1. \tag{5.15}$$

This extremum is determined by the Euler-Lagrange equations

$$\frac{\partial G}{\partial q_k} - \frac{d}{ds} \frac{\partial G}{\partial \dot{q}_k} = 0 \qquad (k=1,2,\ldots),$$

where $G = G(q,\dot{q})$ is given by (5.14), and the conditions in (5.15).

We obtain

$$2\lambda(s)\, q_k - \ddot{q}_k = 0 \quad (k = 1,2,\ldots) \tag{5.16}$$

However, by the first equation of (5.15)

$$\sum_{k=1}^{\infty} q_k \dot{q}_k = 0,$$

and so

$$\sum_{k=1}^{\infty} \dot{q}_k^2 + \sum_{k=1}^{\infty} q_k \ddot{q}_k = 0,$$

or, by the second equation of (5.15),

$$\sum_{k=1}^{\infty} q_k \ddot{q}_k = -1$$

It follows from (5.16) and (5.15) that $2\lambda(s)r^2 = -1$ and that

$$\ddot{q}_k + \frac{1}{r^2} q_k = 0 \quad (k=1,2,\ldots).$$

This shows that the geodesic curves $q = q[s]$ of ds_2^2 on $L^2(r)$ are the "great circles"

$$q[s] = a \cos \frac{s}{r} + b \sin \frac{s}{r} \quad (0 \leq s \leq L), \tag{5.17}$$

where a and b are parameter-independent orthogonal functions in $L^2(r)$, i.e.

$$\|a\| = \|b\| = r, \ (a,b) = 0. \tag{5.18}$$

In order to find the geodesic distance $S_2(q_1,q_2)$, with respect to ds^2, between q_1 and q_2 of $L^2(r)$, we use (5.17)-(5.18) with $q[0] = q_1$, $q[L] = q_2$ and $L = S_2(q_1,q_2)$. This gives the spherical distance

$$S_2(q_1,q_2) = r \cos^{-1}\{\frac{1}{r^2}(q_1,q_2)\} \ (q_1,q_2 \in L^2(r)). \tag{5.19}$$

The arc on the great circle in (5.17)-(5.18) connecting the two points $q_1,q_2 \in L^2(r)$ admits the alternative representation

$$\{q[s]\}^2 = A \cos^2(B - \frac{s}{r}) \quad (0 \leq s \leq L) \tag{5.20}$$

where

$$A = \{q_1^2 + q_2^2 - 2q_1q_2 \cos \frac{L}{r}\} / \sin^2 \frac{L}{r}, \tag{5.21}$$

$$B = \tan^{-1} \{(\frac{q_2}{q_1} - \cos \frac{L}{r}) \; / \; \sin \frac{L}{r}\} \tag{5.22}$$

and $L = S_2(q_1,q_2)$.

We now describe the geometries of L_+^α and P^α ($\alpha \in \mathbb{R}$) under the metric ds_α^2. This is done, as mentioned previously, by considering the geometries of L^2 and $L^2(r)$ under ds_2^2 and using the embedding T_α in (5.1) with (5.2)-(5.3) and (5.6)-(5.7). Here $r = r_\alpha$ with $r_\alpha = 2/|\alpha|$ for $\alpha \neq 0$ and $r_0 = 1$.

The previous analysis shows that the geometry of L_+^α under ds_α^2 is essentially euclidean. Thus the Riemann-Christoffel tensor of the first kind is identically zero and the geodesic curves in L_+^α have the following description: For $\alpha = 0$, we have

$$p[s] = ba^s \quad (s \in \mathbb{R})$$

where a and b are parameter-independent functions in L_+^0. For $\alpha \neq 0$, on the other hand, we have

$$p[s] = (as + b)^{2/\alpha} \quad (0 \leq s < \infty)$$

where a and b are parameter-independent functions in L_+^2. The geodesic distance is then

$$S_\alpha(p_1,p_2) = \rho_\alpha(p_1,p_2) \quad (p_1,p_2 \in L_+^\alpha),$$

where ρ_α is the distance given in (5.4)-(5.5).

The geometry of P^α under ds_α^2, on the other hand, is induced by the spherical representation of $L^2(r_\alpha)$. Thus the Riemann-Christoffel tensor of the first kind is

$$R_\alpha(x,y: u,v) = R_2(x,y: u,v) \qquad (x,y,u,v \in L^2),$$

where $R_2(x,y: u,v)$ is given by (5.10). It follows from (5.11) that the mean Gaussian curvature of ds_α^2 in P^α is

$$\kappa_\alpha(x,y) \equiv 1/4 \qquad (x,y \in L^2).$$

Note that the above quantities for $\alpha = 1$ give the first information curvature tensor and the information curvature, respectively.

The geodesic curves and distance on P^α with respect to ds_α^2 are determined via (5.17)-(5.22) with $r = r_\alpha$. For $\alpha = 0$, we have $r_0 = 1$, and the geodesic curves $p = p[s] \in P^0$ are given by

$$p(s) = \exp\{a \cos s + b \sin s\} \quad (0 \leq s \leq L)$$

where a and b are parameter-independent orthonormal functions in $L^2(1)$, i.e.

$$\|a\| = \|b\| = 1, \quad (a,b) = 0.$$

Similarly, the geodesic distance on P^0 is then

$$S_0(p_1,p_2) = \cos^{-1}(\log p_1, \log p_2)$$

or

$$S_0(p_1,p_2) = \cos^{-1} \int_X (\log p_1)(\log p_2) d\mu \quad (p_1,p_2 \in P^0).$$

Moreover, corresponding to (5.20)-(5.22) we also have the alternative representation for the geodesic curve $p = p[s] \in P^0$ connecting the points p_1 and p_2 of P^0, namely

$$p[s] = \exp\{A^{\frac{1}{2}} \cos (B-s)\} \qquad (0 \leq s \leq L)$$

where

$$A = \{(\log p_1)^2+(\log p_2)^2 - 2(\log p_1)(\log p_2)\cos L\} \,/\, \sin^2 L,$$

$$B = \tan^{-1}\{(\frac{\log p_2}{\log p_1}) - \cos L)/\sin L\}$$

and $L = S_0(p_1,p_2)$.

For $\alpha \neq 0$, on the other hand, $r_\alpha = 2/|\alpha|$ and the geodesic curves $p = p[s] \in P^\alpha$ are given by

$$p[s] = \{a \cos \frac{|\alpha|}{2} s + b \sin \frac{|\alpha|}{2} s\}^{2/\alpha} \quad (0 \leq s \leq L),$$

where a and b are parameter-independent orthonormal functions in P^2, i.e.

$$\|a\| = \|b\| = 1, \ (a,b) = 0 \qquad (a,b \in M_+).$$

It is also assumed that

$$a \cos \frac{|\alpha|}{2} s + b \sin \frac{|\alpha|}{2} s \in M_+ \quad (0 \leq s \leq L).$$

In a similar fashion the geodesic distance on P^α is

$$S_\alpha(p_1,p_2) = \frac{2}{|\alpha|} \cos^{-1} (p_1^{\alpha/2},p_2^{\alpha/2})$$

or

$$S_\alpha(p_1,p_2) = \frac{2}{|\alpha|} \cos^{-1} \int_X (p_1p_2)^{\alpha/2} d\mu \quad (p_1,p_2 \in P^\alpha).$$

Moreover, in correspondance with (5.20)-(5.22) the geodesic curve $p = p[s] \in P^\alpha$ connecting the p_1 and p_2 of P^α admits the alternative representation

$$p[s] = A^{1/\alpha} \cos^{2/\alpha} (B - \frac{|\alpha|}{2} s) \quad (0 \leq s \leq L)$$

where

$$A = \{p_1^\alpha + p_2^\alpha - 2(p_1p_2)^{\alpha/2} \cos \frac{|\alpha|}{2} L\} \,/\, \sin^2 \frac{|\alpha|}{2} L,$$

$$B = \tan^{-1}\{[(\frac{p_2}{p_1})^{\alpha/2} - \cos \frac{|\alpha|}{2} L] \,/\, \sin \frac{|\alpha|}{2} L\}$$

and $L = S_\alpha(p_1,p_2)$.

When $\alpha = 1$, we find that the Rao distance on $P = P^1$ is

$$S(p_1,p_2) = S_1(p_1,p_2) = 2 \cos^{-1} \int_X (p_1 p_2)^{\frac{1}{2}} d\mu$$

which is effictively the Hellinger-Bhattacharyya distance, described in 4.3(3°). This distance was obtained previously in Rao [24] by using rather concrete and explicit methods, and later in Dawid [12] by using abstract methods.

REFERENCES

[1] Amari, S., Theory of information spaces-a geometrical foundation of the analysis of communication systems, RAAG Memoirs 4 (1968), 373-418.

[2] Amari, S., Theory of information space: a differential-geometrical foundation of statistics, RAAG Reports 106 (1980), 1-53.

[3] Atkinson, C. and Mitchell, A.F.S., Rao's distance measure, Sankhyā 43 (1981), 345-365.

[4] Barndorff-Nielsen, O., Information and Exponential Families in Statistical Theory, Wiley, New York, 1978.

[5] Bhattacharyya, A., On a measure of divergence between two statistical populations, Bull. Calcutta Math. Soc. 35 (1943), 99-109.

[6] Burbea, J., J-divergences and related concepts, Encycl. Statist. Sci. 4 (1983), 290-296 (ed. Kotz-Johnson), J. Wiley, New York, 1983.

[7] Burbea, J. and Rao, C.R., Entropy differential metric, distance and divergence measures in probability spaces: a unified approach, J. Multivariate Anal. 12 (1982), 575-596.

[8] Burbea, J. and Rao, C.R., Differential metrics in probability spaces, Probability Math. Statist. 3 (1982), 115-132.

[9] Ćenčov, N.N., Categories of mathematical statistics (in Russian), Doklady Akad. Nauk, SSSR 164 (1965), 3.

[10] Čenčov, N.N., Statistical Decision Rules and Optimal Conclusions (in Russian), Nauka, Moskva, 1972.

[11] Dawid, A.P., Discussion on Professor Efron's paper (1975), Ann. Statist. 3 (1975), 1231-1234.

[12] Dawid, A.P., Further comments on some comments on a paper by Bradley Efron, Ann. Statist. 5 (1977), 1249.

[13] Efron, B., Defining the curvature of a statistical problem (with applications to second order deficiency), (with discussion), Ann. Statist. 3 (1975), 1189-1217.

[14] Efron, B., The geometry of exponential families, Ann. Statist. 6 (1978), 362-376.

[15] Eisenhart, L., Riemannian Geometry, Princeton Univ. Press, Princeton, 1926 and 1960.

[16] Eisenhart, L., An Introduction to Differential Geometry, Princeton Univ. Press, Princeton, 1940 and 1964.

[17] Fisher, R.A., Theory of statistical estimation, Proc. Camb. Phil. Soc. 22 (1925), 700-725.

[18] Hicks, N.J., Notes on Differential Geometry, Van Nostrand, Princeton, 1965.

[19] Kobayashi, S. and Nomizu, K., Foundations of Differential Geometry, Vol. II, Wiley, New York, 1968.

[20] Laugwitz, D., Differential and Riemannian Geometry, Academic Press, New York, 1965.

[21] Mahalanobis, P.C., On the generalized distance in statistics, Proc. Nat. Inst. Sci. India 12 (1936), 49-55.

[22] Oller, J.M. and Cuadras, C.M., Rao's distance for negative multinomial distributions, Sankhyā (in press).

[23] Pitman, E.J.G., Some Basic Theory for Statistical Inference, Halsted Press, New York, 1979.

[24] Rao, C.R., Information and accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc. 37 (1945), 81-91.

[25] Rao, C.R., On the distance between two populations, Sankhyā 9 (1949), 246-248.

[26] Rao, C.R., Efficient estimates and optimun inference procedures in large samples, (with discussion), J. Roy. Statist. Soc. B. 24 (1962), 46-72.

[27] Rao, C.R., Linear Statistical Inference and its Applications, Wiley, New York, 1973.

[28] Reeds, J., Discussion on Professor Efron's paper (1975), Ann. Statist. 5 (1977), 1234-1238.

[29] Schouten, J.A., Ricci-Calculus, Springer-Verlag, Berlin, 1954.

[30] Siegel, C.L., Symplectic Geometry, Academic Press, New York, 1964.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) 84-52	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 85-0015
6a. NAME OF PERFORMING ORGANIZATION University of Pittsburgh 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research
6c. ADDRESS (City, State and ZIP Code) Center for Multivariate Analysis 515 Thackeray Hall, Pittsburgh PA 15260	7b. ADDRESS (City, State and ZIP Code) Directorate of Mathematical & Information Sciences, Bolling AFB DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR 8b. OFFICE SYMBOL (If applicable) NM	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-85-C-0008
8c. ADDRESS (City, State and ZIP Code) Bolling AFB DC 20332-6448	10. SOURCE OF FUNDING NOS. PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2304; TASK NO. A5; WORK UNIT NO.
11. TITLE (Include Security Classification) INFORMATIVE GEOMETRY OF PROBABILITY SPACES	
12. PERSONAL AUTHOR(S) Jacob Burbea	
13a. TYPE OF REPORT Technical 13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Yr., Mo., Day) DEC 84 15. PAGE COUNT 55
16. SUPPLEMENTARY NOTATION	
17. COSATI CODES FIELD / GROUP / SUB. GR.	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The paper is concerned with the geometrical properties that are induced by the local information contents and structures of the parameter space of probability distributions. Of particular interest in this investigation is the Rao distance which is the geodesic distance induced by the differential metric associated with the Fisher information matrix of the parameter space. Moreover, following Efron, Dawid and Amari, some affine connections are introduced into the informative geometry of parameter space and thereby elucidating the role of the curvature in statistical studies. In addition, closed form expressions of the Rao distances for certain families of probability distributions are given and discussed.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL MAJ Brian W. Woodruff	22b. TELEPHONE NUMBER (Include Area Code) (202) 767- 5027 22c. OFFICE SYMBOL NM